

Solution 1: Entropy

A fair die is rolled at the same time as a fair coin is tossed. Let A be the number on the upper surface of the die and let B describe the outcome of the coin toss, where

$$B = \begin{cases} 1, & \text{head,} \\ 0, & \text{tail.} \end{cases}$$

Two random variables X and Y are given by $X = A + B$ and $Y = A - B$, respectively.

- (a) Calculate the entropies $H(X)$ and $H(Y)$, the conditional entropies $H(Y|X)$ and $H(X|Y)$, the joint entropy $H(X, Y)$ and the mutual information $I(X; Y)$.

Solution:

Let a, b, x , and y denote the realisations of the random variables A, B, X , and Y , respectively. Each event (a, b) is associated with exactly one event (x, y) and the probability for such an event is given by

$$p_{AB}(a, b) = p_{XY}(x, y) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

Consequently, we obtain for the joint entropy

$$\begin{aligned} H(X, Y) &= - \sum_{x, y} p_{X, Y}(x, y) \log_2 p_{X, Y}(x, y) = -12 \cdot \frac{1}{12} \log_2 \frac{1}{12} \\ &= \log_2 12 \\ &= 2 + \log_2 3 \end{aligned}$$

Below we list the possible values of the random variables X and Y , the associated events (a, b) , and the probability masses $p_X(x)$ and $p_Y(y)$.

x	events (a, b)	$p_X(x)$	y	events (a, b)	$p_Y(y)$
1	(1, 0)	1/12	0	(1, 1)	1/12
2	(2, 0), (1, 1)	1/6	1	(1, 0), (2, 1)	1/6
3	(3, 0), (2, 1)	1/6	2	(2, 0), (3, 1)	1/6
4	(4, 0), (3, 1)	1/6	3	(3, 0), (4, 1)	1/6
5	(5, 0), (4, 1)	1/6	4	(4, 0), (5, 1)	1/6
6	(6, 0), (5, 1)	1/6	5	(5, 0), (6, 1)	1/6
7	(6, 1)	1/12	6	(6, 0)	1/12

The random variable $X = A + B$ can take the values 1 to 7. The probability masses $p_X(x)$ for the values 1 and 7 are equal to 1/12, since they correspond to exactly one event. The probability masses for the values 2 to 6 are equal to 1/6, since each of these values corresponds to two events (a, b) . An analogue result is obtained for the random variable $Y = A - B$.

The marginal entropies are given by

$$\begin{aligned} H(X) &= - \sum_x p_X(x) \log_2 p_X(x) \\ &= -2 \cdot \frac{1}{12} \log_2 \frac{1}{12} - 5 \cdot \frac{1}{6} \log_2 \frac{1}{6} \\ &= \frac{1}{6} \cdot (\log_2 4 + \log_2 3) + \frac{5}{6} \cdot (\log_2 2 + \log_2 3) \\ &= \frac{7}{6} + \log_2 3 \end{aligned}$$

and for Y

$$\begin{aligned}
 H(Y) &= - \sum_y p_Y(y) \log_2 p_Y(y) \\
 &= -2 \cdot \frac{1}{12} \log_2 \frac{1}{12} - 5 \cdot \frac{1}{6} \log_2 \frac{1}{6} \\
 &= \frac{1}{6} \cdot (\log_2 4 + \log_2 3) + \frac{5}{6} \cdot (\log_2 2 + \log_2 3) \\
 &= \frac{7}{6} + \log_2 3
 \end{aligned}$$

We can determine the conditional entropies using

$$H(X|Y) = H(X, Y) - H(Y) = 2 + \log_2 3 - \frac{7}{6} - \log_2 3 = \frac{5}{6}$$

$$H(Y|X) = H(X, Y) - H(X) = 2 + \log_2 3 - \frac{7}{6} - \log_2 3 = \frac{5}{6}$$

The mutual information $I(X; Y)$ can be determined according to

$$I(X; Y) = H(X) - H(X|Y) = \frac{7}{6} + \log_2 3 - \frac{5}{6} = \frac{1}{3} + \log_2 3$$

or

$$I(X; Y) = H(Y) - H(Y|X) = \frac{7}{6} + \log_2 3 - \frac{5}{6} = \frac{1}{3} + \log_2 3$$

(b) Show that, for independent discrete random variables X and Y ,

$$I(X; X + Y) - I(Y; X + Y) = H(X) - H(Y)$$

Solution:

Using the definition of mutual information for discrete random variables, $I(X; Y) = H(Y) - H(Y|X)$, we can write

$$\begin{aligned}
 I(X; X + Y) - I(Y; X + Y) &= H(X + Y) - H(X + Y|X) - H(X + Y) + H(X + Y|Y) \\
 &= H(X|Y) - H(Y|X) \\
 &= H(X) - H(Y).
 \end{aligned}$$

The first step follows from the fact that modifying the mean of a pmf doesn't change the entropy. For the second step, we used the fact that the conditional entropy $H(X|Y)$ is equal to the marginal entropy $H(X)$ for independent random variables X and Y .

Solution 2: Kullback-Leibler Divergence

(a) Let f be the pmf of the $\text{Bin}(n, p)$ distribution and q the density of the $\mathcal{N}(\mu, \sigma^2)$.

(i)

$$D_{KL}(f||q) = \mathbb{E}_f \left[\log \frac{f(X)}{q(X, \theta)} \right] = \mathbb{E}_f [\log f(X)] - \mathbb{E}_f [\log q(X|\theta)]$$

(ii) For the gradients, we must derive the partial derivatives of the second part of the KLD. The involved log-density is

$$\log q(X|\theta) = \text{const.} - 0.5 \log \sigma^2 - \frac{1}{2\sigma^2} (X - \mu)^2.$$

$$\partial D_{KL}(f||q) / \partial \mu = \partial - \mathbb{E}_f \log[q(X|\theta)] = \mathbb{E}_f \frac{1}{\sigma^2} (X - \mu) \quad (1)$$

$$\partial D_{KL}(f||q)/\partial \sigma^2 = \partial - \mathbb{E}_f \log[q(X|\theta)] = \mathbb{E}_f \left[\frac{1}{2\sigma^2} + \frac{-1}{2\sigma^4} (X - \mu)^2 \right] \quad (2)$$

(iii) Yes, there is. We can first set (1) to zero and get: $\mu = \mathbb{E}_f(X) \Leftrightarrow \mu = np$. We then use this solution for the second equation (2), which we also set to zero first:

$$(2) = 0 \Leftrightarrow \sigma^2 = \mathbb{E}_f[(X - \mu)^2] = \text{Var}_f(X) + (\mathbb{E}_f[X - \mu])^2 = np(1 - p) + (\mathbb{E}_f[X - \mu])^2.$$

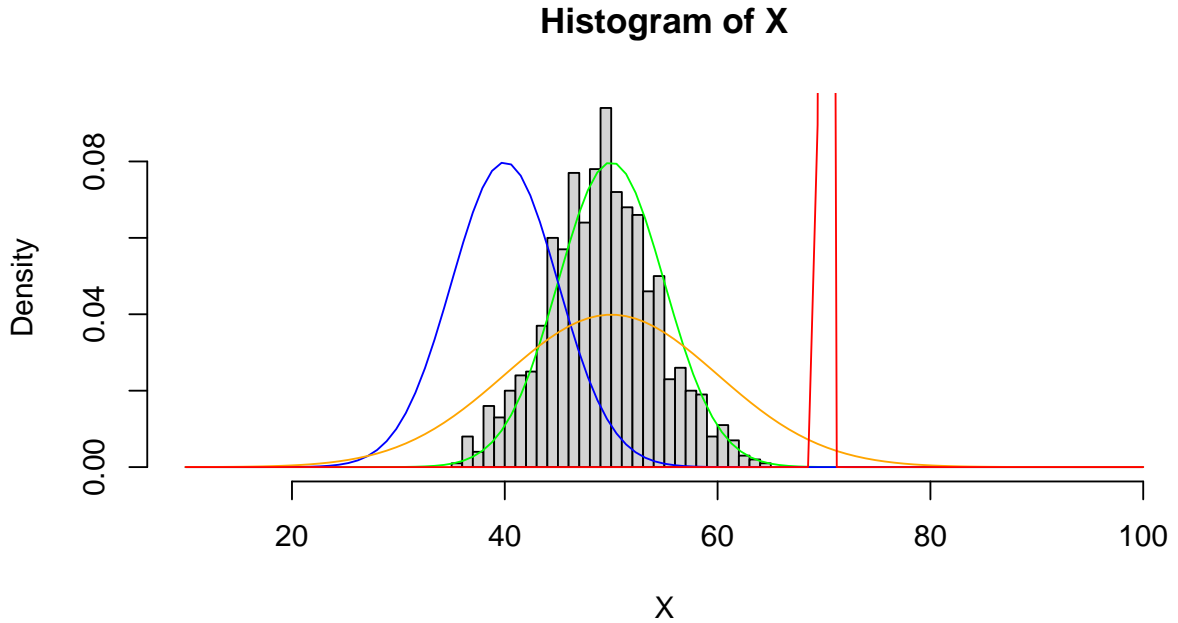
Using $\mu = np$, the second term vanishes and we get the optimal $\sigma^2 = np(1 - p) = \text{Var}_f(X)$. Note that we would have to prove that the second derivative is < 0 to be sure that we found a minimum!

(iv) We could, alternatively, use the gradients and do gradient descent to find the optimal θ .

```
(b) nr_points = 1000
p = 0.5
n = 100
# create data
X <- rbinom(nr_points, prob = p, size = n)

# define different Normal density functions
normal_optimal <- function(x) dnorm(x, mean = n*p, sd = sqrt(n*p*(1-p)))
normal_shift <- function(x) dnorm(x, mean = n*p - 10, sd = sqrt(n*p*(1-p)))
normal_scale_increase <- function(x) dnorm(x, mean = n*p, sd = sqrt(n*p*(1-p))*2)
normal_right_scale_decrease <- function(x) dnorm(x, mean = n*p + 20, sd = p*(1-p))

hist(X, breaks = 25, xlim = c(10, 100), freq = FALSE)
curve(normal_optimal, from = 10, to = 100, add = TRUE, col = "green")
curve(normal_shift, from = 10, to = 100, add = TRUE, col = "blue")
curve(normal_scale_increase, from = 10, to = 100, add = TRUE, col = "orange")
curve(normal_right_scale_decrease, from = 10, to = 100, add = TRUE, col = "red")
```



For these distributions, we get the following KL divergence values (up to an additive constant):

$$D_{KL}(f||q) = \text{const.} + 0.5 \log \sigma^2 + \frac{1}{2\sigma^2} (\text{Var}_f(X) + (np - \mu)^2)$$

```

kld_value <- function(mu,sigma2)
{
  0.5*log(sigma2) +
  0.5 * (sigma2)^(-1) * (n*p*(1-p) + (n*p - mu)^2)
}
(optimal_green <- kld_value(n*p,n*p*(1-p)))

## [1] 2.109438

(shift_blue <- kld_value(n*p-10,n*p*(1-p)))

## [1] 4.109438

(scale_increase_orange <- kld_value(n*p,n*p*(1-p)*4))

## [1] 2.427585

(right_scale_decrease_red <- kld_value(n*p+20, (p*(1-p))^2))

## [1] 3398.614

```

- (c) Since we are now required to calculate the exact KLD values, we would also have to calculate $\mathbb{E}_f(\log f(X))$, which is somewhat more difficult. If you search the internet for a solution (\rightarrow “entropy of a binomial distribution”), you will find an approximate solution using the de-Moivre-Laplace theorem. Alternatively, we could make use of the central limit theorem, but then we would just approximate f with a normal distribution with $\mu = np$ and $\sigma^2 = np(1 - p)$, which would give us a constant KLD of zero (the very same happens if you use the first approximation using the de-Moivre-Laplace-theorem). We here instead will approximate the expectation using a large sample from the true underlying distribution:

$$D_{KL}(f||q) \approx \frac{1}{B} \sum_{b=1}^B [\log f(X) - \log q(X|\mu = np, \sigma^2 = np(1 - p))]$$

```

p_seq <- seq(0.01, 0.99, l = 100)
n_seq <- seq(10, 500, by = 100)
B <- 10000

kld_value_approx <- function(n,p){

  # sample a large number of data points from true distribution
  x <- rbinom(B, prob = p, size = n)

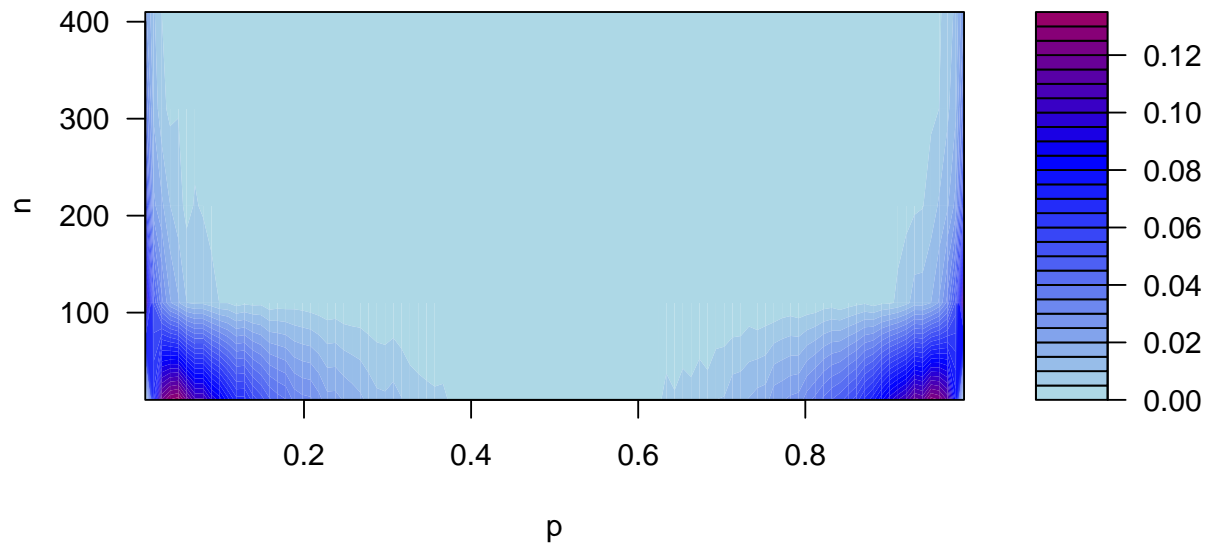
  # approximate the mean; threshold values to 0 if < 0 due
  # to the approximation
  pmax(
    mean(
      dbinom(x, prob = p, size = n, log = TRUE) -
      dnorm(x, mean = n*p, sd = sqrt(n*p*(1-p)), log = TRUE),
      na.rm = TRUE
    ),
    0
  )
}

kld_val <- sapply(n_seq, function(this_n)
  sapply(p_seq, function(this_p) kld_value_approx(this_n, this_p)))

```

```
cols = rev(colorRampPalette(c('darkred', 'red', 'blue', 'lightblue'))(50))

filled.contour(x = p_seq, y = n_seq, z = kld_val,
               xlab = "p", ylab = "n",
               col = cols
               )
```



- (d) Based on the previous result, one can see that the KLD is very close to zero but has larger values for very small or very large values of p and n in combination with a small number of experiments n . These are exactly the cases where the normal approximation of a binomial distribution does not work so well.